

10-10-00

IN THE U.S. PATENT AND TRADEMARK OFFICE
Patent Application Transmittal Letter

ASSISTANT COMMISSIONER FOR PATENTS
Washington, D.C. 20231

To:

Transmitted herewith for filing under 37 CFR 1.53(b) is a(n): Utility Design

original patent application,
 continuation-in-part application

INVENTOR(S): Bin Zhang, et al.

TITLE: Aggregated Clustering Method and System

Enclosed are:

The Declaration and Power of Attorney. signed unsigned or partially signed
 4 sheets of drawings (one set) Associate Power of Attorney
 Form PTO-1449 Information Disclosure Statement and Form PTO-1449
 Priority document(s) (Other) (fee \$ _____)

CLAIMS AS FILED BY OTHER THAN A SMALL ENTITY				
(1) FOR	(2) NUMBER FILED	(3) NUMBER EXTRA	(4) RATE	(5) TOTALS
TOTAL CLAIMS	15 — 20	0	X \$18	\$ 0
INDEPENDENT CLAIMS	2 — 3	0	X \$78	\$ 0
ANY MULTIPLE DEPENDENT CLAIMS	0		\$260	\$ 0
BASIC FEE: Design \$310.00 ; Utility \$690.00				\$ 690
TOTAL FILING FEE				\$ 690
OTHER FEES				\$
TOTAL CHARGES TO DEPOSIT ACCOUNT				\$ 690

Charge \$ 690 to Deposit Account 08-2025. At any time during the pendency of this application, please charge any fees required or credit any over payment to Deposit Account 08-2025 pursuant to 37 CFR 1.25. Additionally please charge any fees to Deposit Account 08-2025 under 37 CFR 1.16, 1.17, 1.19, 1.20 and 1.21. A duplicate copy of this sheet is enclosed.

"Express Mail" label no. EL188087829US

Date of Deposit 10/4/00

Respectfully submitted,

Bin Zhang, et al.

By Thomas X. Li

Thomas X. Li

Attorney/Agent for Applicant(s)
Reg. No. 37,079

Date: 10/4/00

Telephone No.: (650) 857-5972

United States Patent Application
For

AGGREGATED CLUSTERING METHOD AND SYSTEM

Inventors:

Bin Zhang
Igor Kleyner
Meichun Hsu

AGGREGATED CLUSTERING METHOD AND SYSTEM

Field of the invention

5 The present invention relates generally to data clustering and more specifically to a method and system for aggregated data clustering.

Background of the Invention

10 Data clustering operates to group or partition a plurality of data points into a predetermined number of clusters or categories based on one or more attributes or features. The efficiency of a clustering algorithm depends on several factors. First, the computation resources required to implement the clustering algorithm is an important consideration. It is generally desirable to reduce the time needed to generate results (often referred to as the convergence rate) and also reduce the amount of computer resources needed to implement the clustering algorithm.

15 Furthermore, as explained in greater detail hereinafter, the prior art methods do not have a very efficient convergence rate.

20 Second, the quality of the generated clusters or categories (often referred to as the convergence quality) is also another important consideration. Ideally, there is one center point for each category or cluster. Unfortunately, the prior art methods often generate clusters or categories with more than one center. These centers are referred to as "trapped centers" (i.e., these centers are trapped by the local data, but actually belong to another cluster or category).

There are many practical and useful applications that can utilize data clustering to improve results. Consequently, there is much interest in developing clustering algorithms or methods that efficiently and effectively cluster data.

PRIOR ART DATA CLUSTERING METHODS

K-Means is a well-known prior art method for data clustering. The K-Means clustering algorithm is further described in J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," pages 281-297 in: L. M. Le Cam & J. Neyman [eds.] Proceedings 5 of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press, Berkeley, 1967 and Shokri Z. Selim and M. A. Ismail, "K-Means Type of Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-6, No.1, 1984. Unfortunately, both of these approaches are limited to moving a single data point at one time 10 from one cluster to a second cluster.

15 Data points do not "move" in the physical sense, but each data point's membership in a particular cluster, which is defined by a center point, changes. For example, when a data point that is a member in a first cluster is "moved" to a second cluster, a performance function is evaluated based on the data points and center points before and after the move. One aspect of the clustering algorithm is to determine whether such a "move" reduces the performance function (i.e., whether the "move" improves the clustering results).

20 It is to be appreciated that moving one data point at a time between two clusters is inefficient especially when many thousands, tens of thousands of data points, or more need to be moved. One can analogize this situation with a more common example of negotiating the best price for an automobile.

Consider an example when a seller and a buyer are separated by a difference of five thousand dollars between an initial offer price (e.g., \$10,000) and a counter offer price (e.g., \$15,000). During this stage of the negotiations, it would be very inefficient if the buyer's second offer is \$10,000.01 and the seller counters with \$14,999.99. In fact, if the negotiations were to 5 continue one cent at a time, it is apparent that both the seller and buyer would be negotiating for a long time to come before reaching any type of agreement. Consequently, the speed at which an agreement is reached one cent at a time is very slow at best.

Instead, it would be more efficient, and one would expect the buyer in real life to move perhaps by a thousand dollars or more in the second offer by offering, for example, \$11,000. Similarly, one would expect the seller to move perhaps a thousand dollars in a counter offer by countering with \$14,000. Perhaps, when the buyer and seller were only one thousand dollars apart, the buyer and seller would then start negotiating in increments of hundreds of dollars. Similarly, when the buyer and seller were only one hundred dollars apart from reaching an agreement, both would begin to negotiate in increments of single dollars and then in cents.

15 The inefficient negotiation strategy of moving one cent at a time, regardless of how far apart the parties are, is comparable to what is currently being performed by prior art clustering methods. Since prior art methods are limited to moving a single data point per iteration, this is similar to negotiating on a per penny basis when in fact the parties (e.g., data points and center points) are thousands of dollars apart.

20 From the above example, it can be appreciated that a mechanism to move more than one data point at a time is desirable. Unfortunately, there is no mechanism for moving more than one data point at a time without losing precision. In fact, if the prior art approaches were to move more than one point at a time, there is no method that exists to quantify the amount of error

injected by moving more than one point at a time.

Accordingly, there remains a need for a method and system for data clustering that can move more than one data point at a time without the loss of precision and that overcomes the disadvantages set forth previously.

SUMMARY OF THE INVENTION

It is an object of the present invention to provide a clustering method and system that is capable of simultaneously moving more than one data point from a first cluster to a second cluster.

5 It is a further object of the present invention to provide a clustering method and system for moving more than one data point at a time from a first cluster to a second cluster while preserving the monotone convergence property (i.e., the property that the performance function decreases after every move that is made of data points between two clusters).

10 It is a further object of the present invention to provide a clustering method and system that provides a predetermined metric for evaluating the move of more than one data point between two clusters, where the predetermined metric includes the geometric center of the set of data points currently being evaluated for move.

15 It is yet another object of the present invention to provide a clustering method and system that provides a procedure for updating the performance function without losing precision or using approximations.

20 An aggregated data clustering method and system. First, the data points to be clustered and a size parameter are received. The size parameter specifies the number of data points to be moved at one time in the clustering algorithm. Next, the data points are clustered by using an aggregated clustering algorithm (e.g., aggregated local K-Means clustering algorithm) and the size parameter to generate clustered results. Then, a determination is made whether or not the clustered results are satisfactory. If the clustered results are satisfactory, the clustering is stopped. Otherwise, a modified or refined parameter size is received. For example, a user can decrease the parameter size to reduce the number of data points that are moved from a first

10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
160
165
170
175
180
185
190
195
200
205
210
215
220
225
230
235
240
245
250
255
260
265
270
275
280
285
290
295
300
305
310
315
320
325
330
335
340
345
350
355
360
365
370
375
380
385
390
395
400
405
410
415
420
425
430
435
440
445
450
455
460
465
470
475
480
485
490
495
500
505
510
515
520
525
530
535
540
545
550
555
560
565
570
575
580
585
590
595
600
605
610
615
620
625
630
635
640
645
650
655
660
665
670
675
680
685
690
695
700
705
710
715
720
725
730
735
740
745
750
755
760
765
770
775
780
785
790
795
800
805
810
815
820
825
830
835
840
845
850
855
860
865
870
875
880
885
890
895
900
905
910
915
920
925
930
935
940
945
950
955
960
965
970
975
980
985
990
995
1000
1005
1010
1015
1020
1025
1030
1035
1040
1045
1050
1055
1060
1065
1070
1075
1080
1085
1090
1095
1100
1105
1110
1115
1120
1125
1130
1135
1140
1145
1150
1155
1160
1165
1170
1175
1180
1185
1190
1195
1200
1205
1210
1215
1220
1225
1230
1235
1240
1245
1250
1255
1260
1265
1270
1275
1280
1285
1290
1295
1300
1305
1310
1315
1320
1325
1330
1335
1340
1345
1350
1355
1360
1365
1370
1375
1380
1385
1390
1395
1400
1405
1410
1415
1420
1425
1430
1435
1440
1445
1450
1455
1460
1465
1470
1475
1480
1485
1490
1495
1500
1505
1510
1515
1520
1525
1530
1535
1540
1545
1550
1555
1560
1565
1570
1575
1580
1585
1590
1595
1600
1605
1610
1615
1620
1625
1630
1635
1640
1645
1650
1655
1660
1665
1670
1675
1680
1685
1690
1695
1700
1705
1710
1715
1720
1725
1730
1735
1740
1745
1750
1755
1760
1765
1770
1775
1780
1785
1790
1795
1800
1805
1810
1815
1820
1825
1830
1835
1840
1845
1850
1855
1860
1865
1870
1875
1880
1885
1890
1895
1900
1905
1910
1915
1920
1925
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
2030
2035
2040
2045
2050
2055
2060
2065
2070
2075
2080
2085
2090
2095
2100
2105
2110
2115
2120
2125
2130
2135
2140
2145
2150
2155
2160
2165
2170
2175
2180
2185
2190
2195
2200
2205
2210
2215
2220
2225
2230
2235
2240
2245
2250
2255
2260
2265
2270
2275
2280
2285
2290
2295
2300
2305
2310
2315
2320
2325
2330
2335
2340
2345
2350
2355
2360
2365
2370
2375
2380
2385
2390
2395
2400
2405
2410
2415
2420
2425
2430
2435
2440
2445
2450
2455
2460
2465
2470
2475
2480
2485
2490
2495
2500
2505
2510
2515
2520
2525
2530
2535
2540
2545
2550
2555
2560
2565
2570
2575
2580
2585
2590
2595
2600
2605
2610
2615
2620
2625
2630
2635
2640
2645
2650
2655
2660
2665
2670
2675
2680
2685
2690
2695
2700
2705
2710
2715
2720
2725
2730
2735
2740
2745
2750
2755
2760
2765
2770
2775
2780
2785
2790
2795
2800
2805
2810
2815
2820
2825
2830
2835
2840
2845
2850
2855
2860
2865
2870
2875
2880
2885
2890
2895
2900
2905
2910
2915
2920
2925
2930
2935
2940
2945
2950
2955
2960
2965
2970
2975
2980
2985
2990
2995
3000
3005
3010
3015
3020
3025
3030
3035
3040
3045
3050
3055
3060
3065
3070
3075
3080
3085
3090
3095
3100
3105
3110
3115
3120
3125
3130
3135
3140
3145
3150
3155
3160
3165
3170
3175
3180
3185
3190
3195
3200
3205
3210
3215
3220
3225
3230
3235
3240
3245
3250
3255
3260
3265
3270
3275
3280
3285
3290
3295
3300
3305
3310
3315
3320
3325
3330
3335
3340
3345
3350
3355
3360
3365
3370
3375
3380
3385
3390
3395
3400
3405
3410
3415
3420
3425
3430
3435
3440
3445
3450
3455
3460
3465
3470
3475
3480
3485
3490
3495
3500
3505
3510
3515
3520
3525
3530
3535
3540
3545
3550
3555
3560
3565
3570
3575
3580
3585
3590
3595
3600
3605
3610
3615
3620
3625
3630
3635
3640
3645
3650
3655
3660
3665
3670
3675
3680
3685
3690
3695
3700
3705
3710
3715
3720
3725
3730
3735
3740
3745
3750
3755
3760
3765
3770
3775
3780
3785
3790
3795
3800
3805
3810
3815
3820
3825
3830
3835
3840
3845
3850
3855
3860
3865
3870
3875
3880
3885
3890
3895
3900
3905
3910
3915
3920
3925
3930
3935
3940
3945
3950
3955
3960
3965
3970
3975
3980
3985
3990
3995
4000
4005
4010
4015
4020
4025
4030
4035
4040
4045
4050
4055
4060
4065
4070
4075
4080
4085
4090
4095
4100
4105
4110
4115
4120
4125
4130
4135
4140
4145
4150
4155
4160
4165
4170
4175
4180
4185
4190
4195
4200
4205
4210
4215
4220
4225
4230
4235
4240
4245
4250
4255
4260
4265
4270
4275
4280
4285
4290
4295
4300
4305
4310
4315
4320
4325
4330
4335
4340
4345
4350
4355
4360
4365
4370
4375
4380
4385
4390
4395
4400
4405
4410
4415
4420
4425
4430
4435
4440
4445
4450
4455
4460
4465
4470
4475
4480
4485
4490
4495
4500
4505
4510
4515
4520
4525
4530
4535
4540
4545
4550
4555
4560
4565
4570
4575
4580
4585
4590
4595
4600
4605
4610
4615
4620
4625
4630
4635
4640
4645
4650
4655
4660
4665
4670
4675
4680
4685
4690
4695
4700
4705
4710
4715
4720
4725
4730
4735
4740
4745
4750
4755
4760
4765
4770
4775
4780
4785
4790
4795
4800
4805
4810
4815
4820
4825
4830
4835
4840
4845
4850
4855
4860
4865
4870
4875
4880
4885
4890
4895
4900
4905
4910
4915
4920
4925
4930
4935
4940
4945
4950
4955
4960
4965
4970
4975
4980
4985
4990
4995
5000
5005
5010
5015
5020
5025
5030
5035
5040
5045
5050
5055
5060
5065
5070
5075
5080
5085
5090
5095
5100
5105
5110
5115
5120
5125
5130
5135
5140
5145
5150
5155
5160
5165
5170
5175
5180
5185
5190
5195
5200
5205
5210
5215
5220
5225
5230
5235
5240
5245
5250
5255
5260
5265
5270
5275
5280
5285
5290
5295
5300
5305
5310
5315
5320
5325
5330
5335
5340
5345
5350
5355
5360
5365
5370
5375
5380
5385
5390
5395
5400
5405
5410
5415
5420
5425
5430
5435
5440
5445
5450
5455
5460
5465
5470
5475
5480
5485
5490
5495
5500
5505
5510
5515
5520
5525
5530
5535
5540
5545
5550
5555
5560
5565
5570
5575
5580
5585
5590
5595
5600
5605
5610
5615
5620
5625
5630
5635
5640
5645
5650
5655
5660
5665
5670
5675
5680
5685
5690
5695
5700
5705
5710
5715
5720
5725
5730
5735
5740
5745
5750
5755
5760
5765
5770
5775
5780
5785
5790
5795
5800
5805
5810
5815
5820
5825
5830
5835
5840
5845
5850
5855
5860
5865
5870
5875
5880
5885
5890
5895
5900
5905
5910
5915
5920
5925
5930
5935
5940
5945
5950
5955
5960
5965
5970
5975
5980
5985
5990
5995
6000
6005
6010
6015
6020
6025
6030
6035
6040
6045
6050
6055
6060
6065
6070
6075
6080
6085
6090
6095
6100
6105
6110
6115
6120
6125
6130
6135
6140
6145
6150
6155
6160
6165
6170
6175
6180
6185
6190
6195
6200
6205
6210
6215
6220
6225
6230
6235
6240
6245
6250
6255
6260
6265
6270
6275
6280
6285
6290
6295
6300
6305
6310
6315
6320
6325
6330
6335
6340
6345
6350
6355
6360
6365
6370
6375
6380
6385
6390
6395
6400
6405
6410
6415
6420
6425
6430
6435
6440
6445
6450
6455
6460
6465
6470
6475
6480
6485
6490
6495
6500
6505
6510
6515
6520
6525
6530
6535
6540
6545
6550
6555
6560
6565
6570
6575
6580
6585
6590
6595
6600
6605
6610
6615
6620
6625
6630
6635
6640
6645
6650
6655
6660
6665
6670
6675
6680
6685
6690
6695
6700
6705
6710
6715
6720
6725
6730
6735
6740
6745
6750
6755
6760
6765
6770
6775
6780
6785
6790
6795
6800
6805
6810
6815
6820
6825
6830
6835
6840
6845
6850
6855
6860
6865
6870
6875
6880
6885
6890
6895
6900
6905
6910
6915
6920
6925
6930
6935
6940
6945
6950
6955
6960
6965
6970
6975
6980
6985
6990
6995
7000
7005
7010
7015
7020
7025
7030
7035
7040
7045
7050
7055
7060
7065
7070
7075
7080
7085
7090
7095
7100
7105
7110
7115
7120
7125
7130
7135
7140
7145
7150
7155
7160
7165
7170
7175
7180
7185
7190
7195
7200
7205
7210
7215
7220
7225
7230
7235
7240
7245
7250
7255
7260
7265
7270
7275
7280
7285
7290
7295
7300
7305
7310
7315
7320
7325
7330
7335
7340
7345
7350
7355
7360
7365
7370
7375
7380
7385
7390
7395
7400
7405
7410
7415
7420
7425
7430
7435
7440
7445
7450
7455
7460
7465
7470
7475
7480
7485
7490
7495
7500
7505
7510
7515
7520
7525
7530
7535
7540
7545
7550
7555
7560
7565
7570
7575
7580
7585
7590
7595
7600
7605
7610
7615
7620
7625
7630
7635
7640
7645
7650
7655
7660
7665
7670
7675
7680
7685
7690
7695
7700
7705
7710
7715
7720
7725
7730
7735
7740
7745
7750
7755
7760
7765
7770
7775
7780
7785
7790
7795
7800
7805
7810
7815
7820
7825
7830
7835
7840
7845
7850
7855
7860
7865
7870
7875
7880
7885
7890
7895
7900
7905
7910
7915
7920
7925
7930
7935
7940
7945
7950
7955
7960
7965
7970
7975
7980
7985
7990
7995
8000
8005
8010
8015
8020
8025
8030
8035
8040
8045
8050
8055
8060
8065
8070
8075
8080
8085
8090
8095
8100
8105
8110
8115
8120
8125
8130
8135
8140
8145
8150
8155
8160
8165
8170
8175
8180
8185
8190
8195
8200
8205
8210
8215
8220
8225
8230
8235
8240
8245
8250
8255
8260
8265
8270
8275
8280
8285
8290
8295
8300
8305
8310
8315
8320
8325
8330
8335
8340
8345
8350
8355
8360
8365
8370
8375
8380
8385
8390
8395
8400
8405
8410
8415
8420
8425
8430
8435
8440
8445
8450
8455
8460
8465
8470
8475
8480
8485
8490
8495
8500
8505
8510
8515
8520
8525
8530
8535
8540
8545
8550
8555
8560
8565
8570
8575
8580
8585
8590
8595
8600
8605
8610
8615
8620
8625
8630
8635
8640
8645
8650
8655
8660
8665
8670
8675
8680
8685
8690
8695
8700
8705
8710
8715
8720
8725
8730
8735
8740
8745
8750
8755
8760
8765
8770
8775
8780
8785
8790
8795
8800
8805
8810
8815
8820
8825
8830
8835
8840
8845
8850
8855
8860
8865
8870
8875
8880
8885
8890
8895
8900
8905
8910
8915
8920
8925
8930
8935
8940
8945
8950
8955
8960
8965
8970
8975
8980
8985
8990
8995
9000
9005
9010
9015
9020
9025
9030
9035
9040
9045
9050
9055
9060
9065
9070
9075
9080
9085
9090
9095
9100
9105
9110
9115
9120
9125
9130
9135
9140

cluster to a second cluster at one time. Then, clustering is performed on the clustered results generated previously by using the aggregated clustering algorithm and the revised or refined parameter size. The steps of determining, modifying the parameter size, and aggregated clustering are repeated until satisfactory clustering results are achieved.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings and in which like reference numerals refer to similar elements.

5 FIG. 1 is an exemplary set of data points that are grouped into a plurality of clusters and that can be the input to the aggregated clustering method of the present invention.

FIG. 2 is a flowchart illustrating an aggregated clustering method according to one embodiment of the present invention.

FIG. 3 is a flowchart illustrating in greater detail certain steps of the flowchart of FIG. 2.

10 FIG. 4 is a block diagram illustration of an aggregated clustering system configured in accordance with one embodiment of the present invention.

USPTO-2015-060000

DETAILED DESCRIPTION

In the following description, for the purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be apparent, however, to one skilled in the art that the present invention may be practiced without 5 these specific details. In other instances, well-known structures and devices are shown in block diagram form in order to avoid unnecessarily obscuring the present invention. The following description and the drawings are illustrative of the invention and are not to be construed as limiting the invention.

DATA CLUSTERING APPLICATION

Before delving into the details of the aggregated clustering method and system of the present invention, an exemplary application is first described to familiarize the reader with concepts related to the invention.

As noted previously, clustering seeks to locate dense regions of data that have similar attributes or features and generate categories or clusters of these "similar" data points. These attributes or features can be a qualitative (e.g., similar behavior, tastes, likes, dis-likes of consumers), or a quantitative measure (e.g., the number of items purchased by customers across a predefined time period).

FIG. 1 is an exemplary set of data points that are grouped into a plurality of clusters that can be the input to the aggregated clustering method of the present invention. As a departure 20 from prior art clustering methods, the aggregated clustering method and system of the present invention moves more than a single data point at one time during the clustering (i.e., changes membership of more than one data point from a first cluster to a second cluster at one time). Specifically, the aggregated clustering method of the present invention can move a plurality of

data points, such as a subset U, from a first cluster (e.g., cluster A) to a second cluster (e.g., cluster B) at one time without the loss of precision. In fact, as explained in greater detail hereinafter, a user can flexibly specify the number of data points in the subset U to be moved at one time by using an input referred to as a parameter size.

5 For example, the set of data points can represent a plurality of car brokers or dealers. This exemplary application uses two attributes or features for the clustering. The first attribute is the number of sedans that the particular dealer has sold in the last year, and the second attribute is the number of sports cars, the particular dealer has sold in the last year.

10 This particular application seeks to group the car dealers into clusters, such as a first cluster (e.g., cluster A) of car dealers that are particularly good at selling sedans, a second cluster (e.g., cluster B) of car dealers that are particularly good at selling sports cars, and perhaps a third cluster (e.g., cluster C) of car dealers that are good at selling both sports cars and sedans.

15 Center-based clustering algorithms operate by receiving the number of desired clusters, initialization information (e.g., the random initial positions of centers), and based thereon generates center points that are at the center of clusters of data. In this case, since there are three desired clusters, three center points with initial points are provided to the clustering algorithm.

20 Ideally, a good clustering method moves the center positions to the three clusters of data (i.e., a first center is moved to the center of those car dealers that sell high numbers of sedans, a second center is moved to the center of those car dealers that sell high numbers of sports cars, and a third center is moved to the center of the car dealers that sell a high number of both sports cars and sedans.

Clustering Method

FIG. 2 is a flowchart illustrating an aggregated clustering method according to one embodiment of the present invention. In step 204, the data points to be clustered and a size parameter are received. The size parameter specifies the number of data points to be moved at one time in the clustering algorithm. In step 208, the data points are clustered using the size parameter to generate clustered results.

In step 214, a determination is made whether or not the clustered results generated in step 208 are satisfactory. A determination of whether results are satisfactory can vary across applications and depend on the specific requirements of the particular application. Typically, one or more well-known metrics is utilized to determine if the clustered results meet a particular requirement. Steps 208 and 214 are described in greater detail hereinafter with reference to FIG. 3.

10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95

When the clustered results are satisfactory, the clustering stops. Otherwise, when the clustered results are not satisfactory, a modified or refined parameter size is received. For example, a user can decrease the parameter size to reduce the number of data points that are moved from a first cluster to a second cluster at one time. By so doing, the granularity of the clustering is increased. One advantage of the present invention is that the user can flexibly select or vary the size parameter to suit a particular clustering application. For example, with a large data set, a user can set the size parameter at a large value such as 1000 for the first iteration, a smaller value, such as 500 for the second iteration, a yet smaller value, such as 100 in a third iteration, etc. In this manner, the aggregated clustering of the present invention allows a user to selectively adjust the granularity of the clustering for each iteration, thereby increasing the efficiency and convergence rate of the clustering.

Furthermore, since the user is not limited to moving a single data point at one time as in the prior art clustering methods, the present invention provide the user the ability to tailor the granularity of the clustering based on the requirements of a particular application.

In step 228, clustering is performed on the clustered results generated by step 208 by 5 using the revised or refined parameter size. Steps 214 through step 228 are repeated until satisfactory clustering results are achieved.

Aggregated Clustering System 400

FIG. 4 illustrates an aggregated clustering system 400 that is configured in accordance with one embodiment of the present invention. The aggregated clustering system 400 includes a move determination unit 404 for evaluating whether an aggregated move of the specified number of data points at one time is possible and enhances the clustering results. The system 400 also includes an aggregated move unit 408 that is coupled to the move determination unit 404 to receive a geometric center 416 of the current set of data points and input information. For example, the move unit 408 updates the first partition count 450, the second partition count 460, 15 the first partition center 454, and the second partition center 464 as described in greater detail hereinafter. Based on these inputs, the move unit 408 accomplishes the move from a first cluster to a second cluster after the move determination unit 404 determines that the aggregated move is needed.

The move determination unit 404 includes a first input for receiving the data points 430 20 that are partitioned into a plurality of initial partitions and a second input for receiving center points 434. As described in greater detail hereinafter, the partitions, center points of the partitions, and the number of data points in each partition (i.e., the count for each partition) may be updated for each iteration of the clustering in accordance with teachings of the present

invention. The move determination unit 404 also includes a third input for receiving the parameter size 438 (i.e., the number of data points to move at one time), a fourth input for receiving information concerning the first partition (i.e., the move from partition) and the second partition (i.e., the move to partition). For example, this information can include the current 5 count 450 of the first partition, the current center 454 of the first partition, the current count 460 of the second partition, and the current center 464 of the second partition.

The move determination unit 404 includes a move evaluation mechanism 412 for evaluating whether a set of data points should be moved from a first cluster to a second cluster. Preferably, the move evaluation mechanism 412 utilizes a predetermined metric 413 for performing the move evaluation. As described in greater detail hereinafter, the predetermined metric 413 can employ a geometric center of the data points considered for move.

The move determination unit 404 also includes a geometric center determination unit 414 for generating the geometric center of the data points to be moved at one time based on the data points in the partitions. As noted previously, the move determination unit 404 uses the geometric center in the move evaluation of a current set of data points. For example, the 15 predetermined metric can include the geometric center of the set of data points evaluated for move. The geometric center of data points is also provided to the aggregated move unit 408 for use in updating the partitions.

The aggregated move unit 408 includes a count update unit 470 for updating the count of 20 the first partition and count of the second partition to accurately reflect the partition counts after the aggregated move. The aggregated move unit 408 also includes a center update unit 474 for updating the center of the first partition and center of the second partition to accurately reflect the partition centers after the aggregated move.

For example, the count update unit 470 adjusts the count for the Move_From partition and the count for the Move_To partition to generate a first partition (FP) count update and second partition (SP) count update, respectively. Specifically, the count for the Move_From partition is decremented by the number of data points involved in the move. Similarly, the count for the Move_To partition is incremented by the number of data points involved in the move. The center update unit 474 adjusts the center of the Move_From partition and the center of the Move_To partition to generate, for example, first partition (FP) center update and second partition (SP) center update, respectively. Specifically, the center for the Move_From partition is calculated by using the geometric center of the data points that are being moved. Similarly, the center for the Move_To partition is calculated by using the geometric center of the data points that are being moved.

These updated partition counts and centers for the first partition and the second partition are then provided to the move determination unit 404 for further move evaluation processing should the current iteration generate clustered results that are not satisfactory.

15 The aggregated clustering method and system of the present invention moves more than one data point at a time from a first cluster to a second cluster while preserving the monotone convergence property. The monotone convergence property is the property that the performance function decreases after every move that is made of data points between two clusters.

Aggregated Clustering

20 FIG. 3 is a flowchart illustrating in greater detail an aggregated clustering method of according to one embodiment of the present invention. In step 304, K initial center positions are received. Each center is denoted by m_k where $k = 1, \dots, K$, where K is the number of clusters (which is herein referred to also as "partitions"). The number of clusters or partitions can be

specified by and adjusted by a user to suit a particular application by setting a size parameter variable. The initial center points can be random points or the output of some initialization algorithm, which are generally known by those of ordinary skill in the art.

In step 308, a plurality of data points are received and partitioned into a plurality of 5 clusters based on the distance of the data point from a center point of a respective cluster. Each cluster has a center point, and each data point is placed into one of the clusters $\{S_k\}$ based on a relationship (e.g., the Euclidean distance) between the center point and the data point.

In step 314, at least two data point in a first partition S_i (e.g., cluster A) are simultaneously evaluated for moving to every other partition (e.g., cluster C and cluster B). For example, subsets of U points are evaluated by an evaluation expression provided herein below. For example, in the example given in FIG. 1, the size parameter is equal to four. All possible combinations or subsets having four data points from the total eleven data points in cluster A are evaluated for move to cluster B or cluster C.

The index i is utilized to represent the partition to which a data point x currently belongs 15 or is a member of, and the index j is utilized to represent the partition that is currently being evaluated for a potential move to which the data point x can be moved. The present invention provides the following predetermined metric for evaluating whether a set of data points should be moved from the current partition to a proposed or potential partition:

$$\frac{n_i}{n_i - |U|} |m_U - m_i|^2 - \frac{n_j}{n_j + |U|} |m_U - m_j|^2$$

where U is the subset of data points (U is a subset of S_i) being evaluated for the move,

20 $|U|$ is the size of U that is specified by the size parameter, m_U is the geometric center of U , m_i

and m_i are the centers of the clusters and n_i and n_j are the counts of the clusters.

In decision block 318, a determination is made whether the value generated in step 314 is greater than zero. When the generated value is greater than zero, processing proceeds to step 344. In step 344, the set of data points U is moved from a current partition S_i to a second partition S_j . Moving the set of data points from a current partition to a second partition can involve the following sub-steps. First, the count of each partition needs to be updated. Second, since the membership of both the partitions are changing (i.e., the data points are being moved from the Move_From partition to the Move_From partition), the centers of these partitions need to be updated and re-calculated. For accomplishing the move U from S_i to S_j , the count of each partition and the center of each partition needs to be re-calculated to accurately reflect the addition of new data points or the deletion of old data points as the case may be.

For updating the counts of the two partitions, the following expressions can be employed:

$$n_i = n_i - |u|, \text{ and} \\ n_j = n_j + |u|.$$

For updating the centers of these two partitions, the following expressions can be employed:

$$m_i = (n_i * m_i - m_u) / (n_i - |u|), \text{ and} \\ m_j = (n_j * m_j + m_u) / (n_j + |u|).$$

If the value generated in step 314 is not greater than zero, processing proceeds to decision block 324, where a determination is made whether there are more data points to be checked. If there are more data points to be checked, then processing proceeds to step 314.

If there are no more data points to be checked, then processing proceeds to decision block 334. Steps 314, 318, 324, and 344 form a single iteration of the processing. In decision block 334, a determination is made whether any data points were moved (i.e., changed

membership in partitions). When no data points are moved (i.e., when no data point changes membership in the partitions), then the processing is complete and stops (step 338). When one or more data points were moved (i.e., at least one data point changed membership in partitions), then processing proceeds to step 314 to process another iteration (i.e., steps 314, 318, 324 and 5 344).

Alternatively, decision block 334 can have a different stop or termination condition. The stop condition can be whether the change in the performance function is less than a predetermined value.

There are numerous applications that can utilize the aggregated clustering method and system of the present invention to cluster data. For example, these applications include, but are not limited to, data mining applications, customer segmentation applications, document categorization applications, scientific data analysis applications, data compression applications, vector quantization applications, and image processing applications.

The foregoing description has provided examples of the present invention. It will be 15 appreciated that various modifications and changes may be made thereto without departing from the broader scope of the invention as set forth in the appended claims.

1 CLAIMS

2 What is claimed is:

3 1. A method for clustering data comprising:
4 (a) receiving a plurality of data points for clustering;
5 (b) receiving a size parameter for specifying the number of data points to be moved at
6 one time;
7 (c) clustering the data points by using the size parameter to generate clustered results;
8 (d) determining whether the clustered results are satisfactory;
9 (e) when the clustered results are satisfactory, stop clustering;
10 (f) otherwise when the clustered results are not satisfactory, revise the size parameter,
11 perform clustering based on the revised size parameter and the clustered results,
12 and proceed to step (d).

13

14 2. The method of claim 1 wherein step (c) further comprises:

15 (c1) evaluating subsets of data points in each cluster for moving into every other cluster
16 by using a predetermined metric; wherein the number of data points in the subset
17 is specified by the size parameter.

18

19 3. The method of claim 2 wherein step (c1) further comprises:

20 (c1_1) determining a geometric center of the subset of data points being evaluated for a
21 move;
22 (c1_2) using the geometric center of the subset of data points in the predetermined metric
23 to generate a value.

24

25 4. The method of claim 3 wherein step (c1) further comprises:

26 (c1_3) determining whether the value is greater than zero;
27 (c1_4) when the value is greater than zero, moving the subset of data points from a
28 Move_From cluster to a Move_To cluster;

29 (c1_5) when the value is not greater than zero, determining if there are more subsets to
30 evaluate;
31 (c1_6) when there are more subsets to evaluate, proceeding to step (c1);
32 (c1_7) when there are no more subsets to evaluate, determining whether any point has
33 moved;
34 (c1_8) when a point has moved, proceeding to step (c1); and
35 (c1_9) when no point has moved, stopping the processing.

36
37 5. The method of claim 4 wherein each data point has a membership with one cluster;
38 wherein step (c1_4) further comprises:

39 simultaneously updating the membership of at least two data points from the membership
40 of the Move_From cluster to the membership of the Move_To cluster.

41
42 6. The method of claim 4 wherein step (c1_4) further comprises:
43 updating the count of the Move_From cluster;
44 updating the center of the Move_From cluster;
45 updating the count of the Move_To cluster; and
46 updating the center of the Move_To cluster.

47
48 7. The method of claim 1 wherein revising the size parameter of step (f) further comprises:
49 (f_1) decreasing the size parameter.

50
51 8. The method of claim 1 wherein step (d) further comprises:
52 (d_1) employing a predetermined metric for determining whether the clustered results are
53 satisfactory; wherein the predetermined metric includes a geometric center of the
54 subset of points that are being evaluated for move.

55
56 9. The method of claim 8 wherein the predetermined metric includes the following
57 expression:

58 where U is the subset of data points being evaluated for the move, $|U|$ is the size of U
59 that is specified by the size parameter, m_{oo} is the geometric center of U , M_I and m_I
60 are the centers of the clusters and n_I and n_j are the counts of the clusters.

61

62 10. The method of claim 1 wherein the clustering method is utilized in one of a data mining
63 application, customer segmentation application, document categorization application, scientific
64 data analysis application, data compression application, vector quantization application, and
65 image processing application.

66

67 11. A clustering system comprising:
68 (a) a source of data points to be clustered; and
69 (b) an aggregated clustering module for moving at least two data points at one time
70 between a Move_From cluster and a Move_To cluster;
71 wherein the aggregated clustering module includes
72 a move determination unit for evaluating the move of subsets of data points from each
73 cluster to every other cluster and determining when such a move should be
74 performed; and
75 an aggregated move unit coupled to the move determination unit that updates a
76 Move_From count, Move_From center, a Move_To count, and a Move_To
77 center.

78

79 12. The clustering system of claim 11 wherein the aggregated clustering module further
80 comprises:
81 (a) a first input for receiving the data points;
82 (b) a second input for receiving initial center points;
83 (c) a third input for receiving a number of points to move at one time;
84 (d) a parameter for storing the center point associated with each cluster
85 (e) a parameter for storing the count of data points associated with each cluster;

86 wherein the data points, center points and counts, are utilized by the move determination
87 unit for move evaluation and determination and by the aggregated move unit for
88 count update and center update.

89

90 13. The clustering system of claim 11 wherein the move determination unit further
91 comprises:

92 a geometric center determination unit for determining the geometric center of a current
93 subset of data points and providing the geometric center to the move
94 determination unit for move evaluation and move determination.

95

96 14. The clustering system of claim 13 wherein the move determination unit further
97 comprises:

98 a move evaluation mechanism for employing a predetermined metric for move
99 evaluation; wherein the predetermined metric includes the geometric center of a
100 current subset of data points.

101
102
103
104
105

102 15. The clustering system of claim 11 is configured for one of a data mining application,
103 customer segmentation application, document categorization application, scientific data analysis
104 application, data compression application, vector quantization application, and image processing
105 application.

106

ABSTRACT OF THE DISCLOSURE

An aggregated data clustering method and system. First, the data points to be clustered and a size parameter are received. The size parameter specifies the number of data points to be moved at one time in the clustering algorithm. Next, the data points are clustered by using an aggregated clustering algorithm (e.g., aggregated local K-Means clustering algorithm) and the size parameter to generate clustered results. Then, a determination is made whether or not the clustered results are satisfactory. If the clustered results are satisfactory, the clustering is stopped. Otherwise, a modified or refined parameter size is received. Then, clustering is performed on the clustered results generated previously by using the aggregated clustering algorithm and the revised or refined parameter size. The steps of determining, modifying the parameter size, and aggregated clustering are repeated until satisfactory clustering results are achieved.

10
10

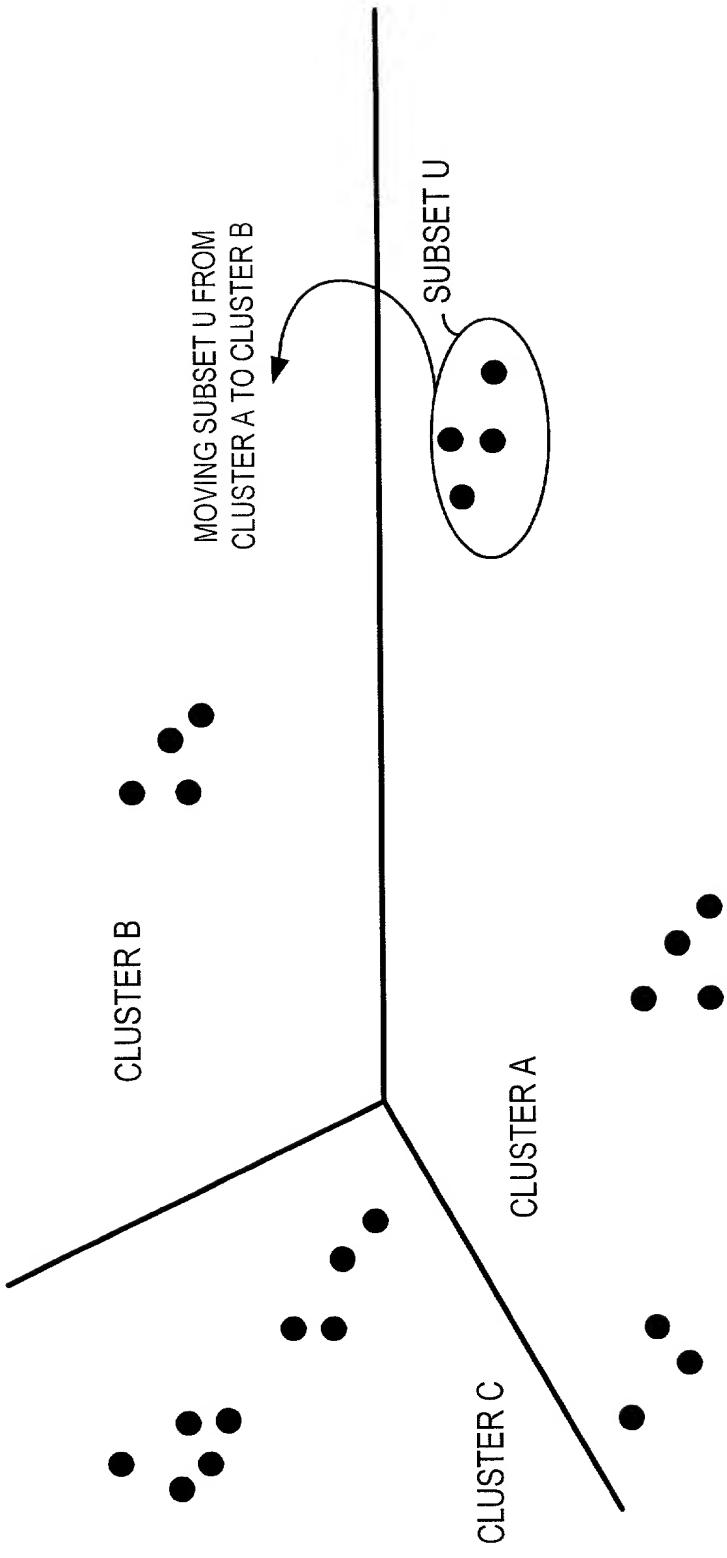


FIG. 1

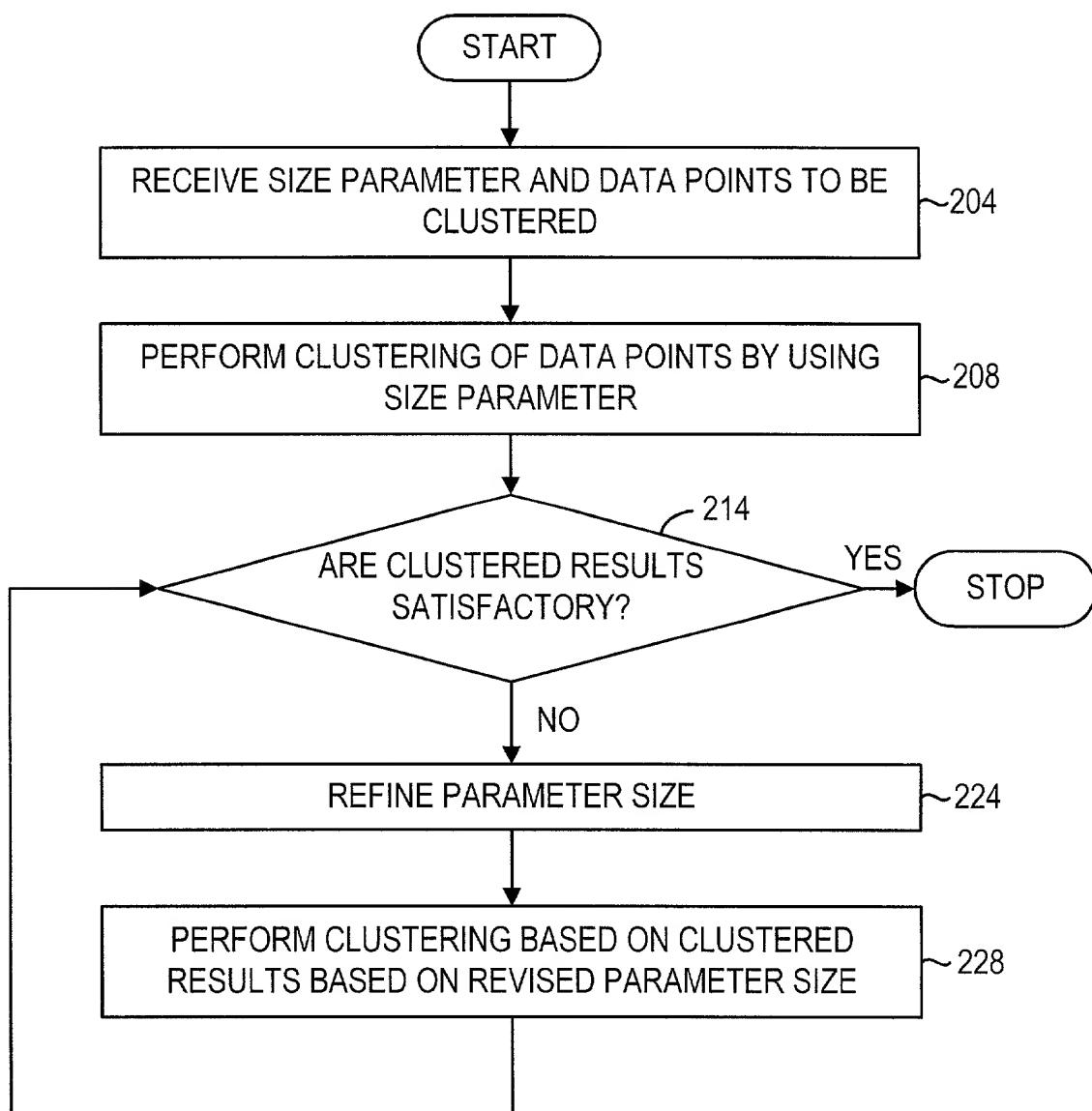


FIG. 2

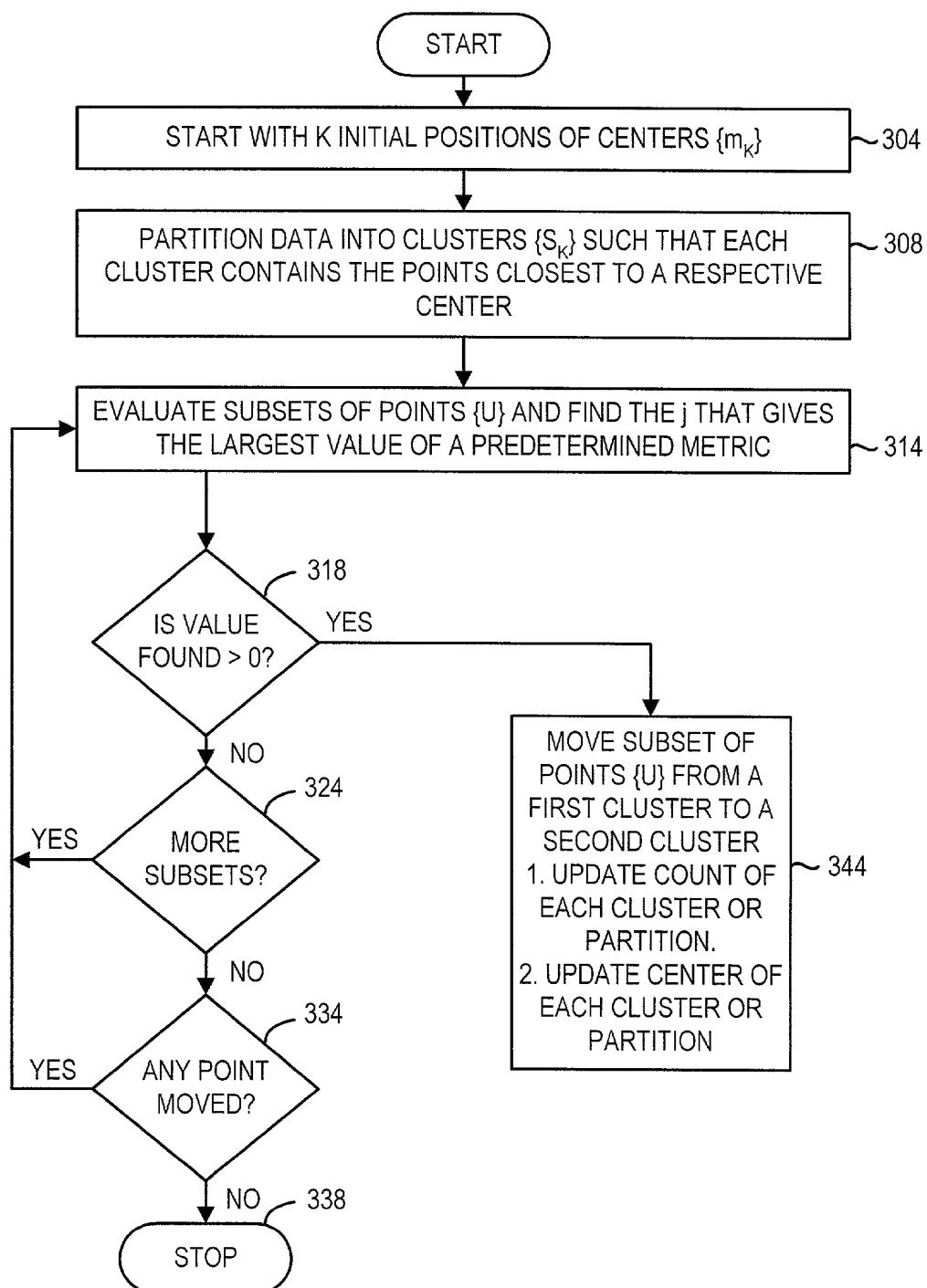
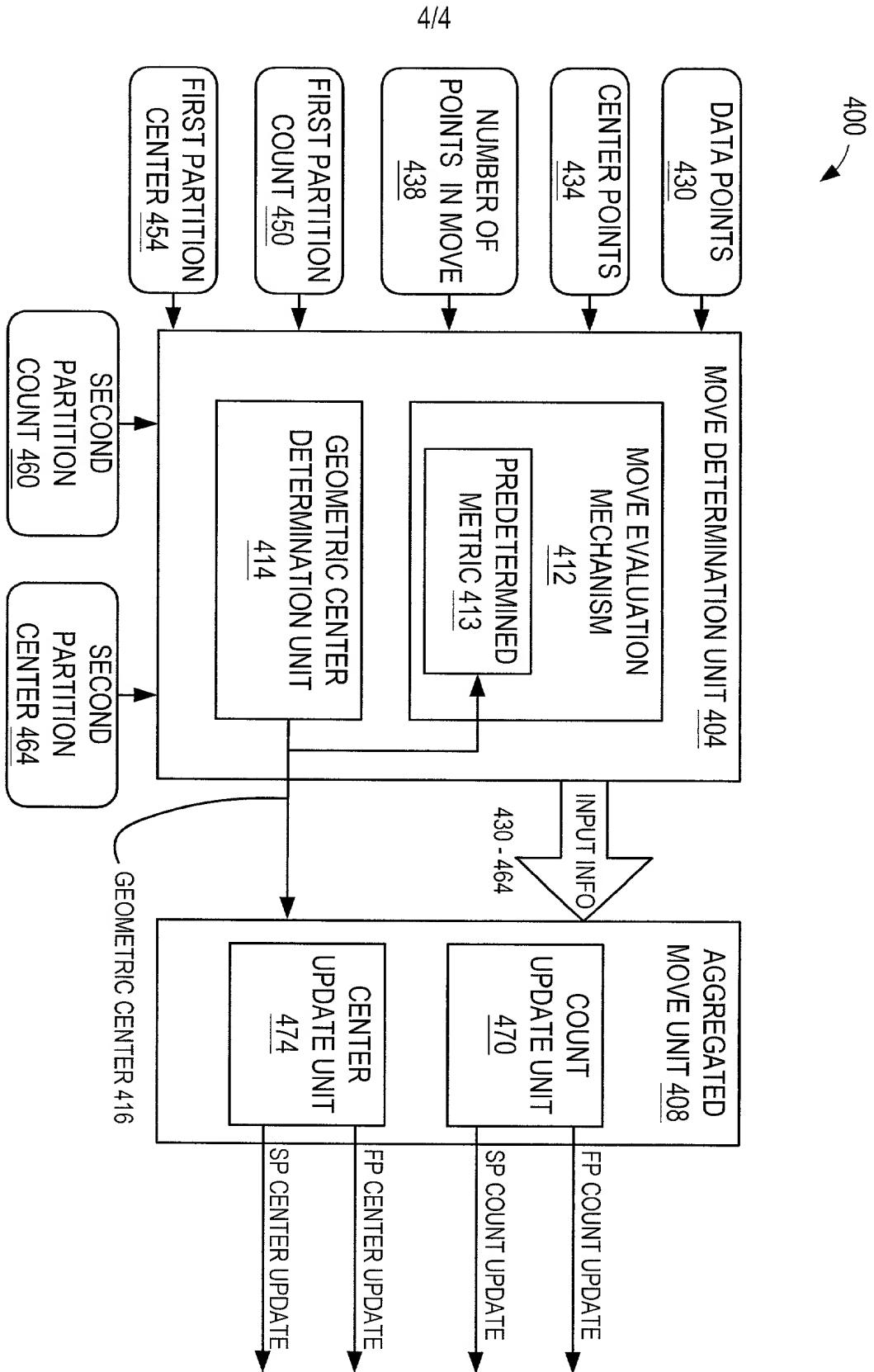


FIG. 3

FIG. 4



**DECLARATION AND POWER OF ATTORNEY
FOR PATENT APPLICATION**

ATTORNEY DOCKET NO. 10992482-1

As a below named inventor, I hereby declare that:

My residence/post office address and citizenship are as stated below next to my name;

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled:

Aggregated Clustering Method and System

the specification of which is attached hereto unless the following box is checked:

was filed on _____ as US Application Serial No. or PCT International Application Number _____ and was amended on _____ (if applicable).

I hereby state that I have reviewed and understood the contents of the above-identified specification, including the claims, as amended by any amendment(s) referred to above. I acknowledge the duty to disclose all information which is material to patentability as defined in 37 CFR 1.56.

Foreign Application(s) and/or Claim of Foreign Priority

I hereby claim foreign priority benefits under Title 35, United States Code Section 119 of any foreign application(s) for patent or inventor(s) certificate listed below and have also identified below any foreign application for patent or inventor(s) certificate having a filing date before that of the application on which priority is claimed:

COUNTRY	APPLICATION NUMBER	DATE FILED	PRIORITY CLAIMED UNDER 35 U.S.C. 119	
N/A			YES: <input type="checkbox"/>	NO: <input type="checkbox"/>
			YES: <input type="checkbox"/>	NO: <input type="checkbox"/>

Provisional Application

I hereby claim the benefit under Title 35, United States Code Section 119(e) of any United States provisional application(s) listed below:

APPLICATION SERIAL NUMBER	FILING DATE
N/A	

U. S. Priority Claim

I hereby claim the benefit under Title 35, United States Code, Section 120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code Section 112, I acknowledge the duty to disclose material information as defined in Title 37, Code of Federal Regulations, Section 1.56(a) which occurred between the filing date of the prior application and the national or PCT international filing date of this application:

APPLICATION SERIAL NUMBER	FILING DATE	STATUS (patented/pending/abandoned)
N/A		

POWER OF ATTORNEY:

As a named inventor, I hereby appoint the following attorney(s) and/or agent(s) to prosecute this application and transact all business in the Patent and Trademark Office connected therewith:

Customer Number **022879**Place Customer
Number Bar Code
Label here

Send Correspondence to:
HEWLETT-PACKARD COMPANY
Intellectual Property Administration
P.O. Box 272400
Fort Collins, Colorado 80528-9599

Direct Telephone Calls To:

Thomas X. Li
(650) 857-5972

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Full Name of Inventor: **Bin Zhang** Citizenship: **Switzerland CHINA**

Residence: **41945 Via San Gabriel, Fremont, CA 94539**

Post Office Address: **Same**

Inventor's Signature _____ Date _____

**DECLARATION AND POWER OF ATTORNEY
FOR PATENT APPLICATION (continued)**

ATTORNEY DOCKET NO. 10992482-1

Full Name of # 2 joint inventor: Meichun Hsu Citizenship: US

Residence: 12717 Leander Drive, Los Altos Hills CA. 94022

Post Office Address: Same

Inventor's Signature _____ Date _____

Full Name of # 3 joint inventor: Igor Kleyner Citizenship: US

Residence: 434 Bally Way, Pacifica, CA 94044

Post Office Address: Same

Inventor's Signature _____ Date _____

Full Name of # 4 joint inventor: _____ Citizenship: _____

Residence: _____

Post Office Address: _____

Inventor's Signature _____ Date _____

Full Name of # 5 joint inventor: _____ Citizenship: _____

Residence: _____

Post Office Address: _____

Inventor's Signature _____ Date _____

Full Name of # 6 joint inventor: _____ Citizenship: _____

Residence: _____

Post Office Address: _____

Inventor's Signature _____ Date _____

Full Name of # 7 joint inventor: _____ Citizenship: _____

Residence: _____

Post Office Address: _____

Inventor's Signature _____ Date _____

Full Name of # 8 joint inventor: _____ Citizenship: _____

Residence: _____

Post Office Address: _____

Inventor's Signature _____ Date _____